ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ & ΑΥΤΟΔΙΟΙΚΗΣΗΣ
ΙΝΣΤΙΤΟΥΤΟ ΕΠΙΜΟΡΦΩΣΗΣ

Στατιστική Συμπερασματολογία με Στατιστικά Πακέτα

Παρουσίαση Εκπαιδευτή

Μαθησιακό Αντικείμενο:

Ανάλυση Παλινδρόμησης και Συσχέτισης

Εκπαιδευτικοί Στόχοι

Με την υλοποίηση του μαθησιακού αντικειμένου, ο καθένας από τους συμμετέχοντες θα μπορεί:

- Να κατανοεί την έννοια της συσχέτισης.
- Να κατανοεί και να ερμηνεύει τα αποτελέσματα της μεθόδου σε απλά και σύνθετα δεδομένα.
- Να εφαρμόζει απλή και πολλαπλή παλινδρόμηση.

Ανάλυση της Συσχέτισης

Ένας εύκολος πρώτος τρόπος για την **ανίχνευση συσχέτισης** μεταξύ *δύο μεταβλητών* είναι το **διάγραμμα διασποράς** (*scatter plot*).

- ✓ **εξαρτημένη μεταβλητή (Y)**
- ✓ **ανεξάρτητη μεταβλητή (X)**

Ανάλυση της Συσχέτισης-Συντελεστής Συσχέτισης (1)

Ο **πληθυσμιακός συντελεστής συσχέτισης** (*population correlation coefficient*) συμβολίζεται με ρ και ορίζεται ως:

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{E(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{E(X_i - \mu_X)^2 (Y_i - \mu_Y)^2}}$$

Αποδεικνύεται ότι $-1 \leq \rho \leq 1$.

Ανάλυση της Συσχέτισης-Συντελεστής Συσχέτισης (2)

Ο **δειγματικός συντελεστής συσχέτισης** (sample correlation coefficient) και ορίζεται ως:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{y})^2}} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}}$$

Παίρνει τιμές στο διάστημα **$-1 \leq r \leq 1$**.

Γραμμική Συσχέτιση (1)

Ο όρος συσχέτιση (correlation) αναφέρεται στον τρόπο που αλληλεπιδρά μια μεταβλητή με μία άλλη. Δύο είναι οι βασικές κατηγορίες συσχέτισης:

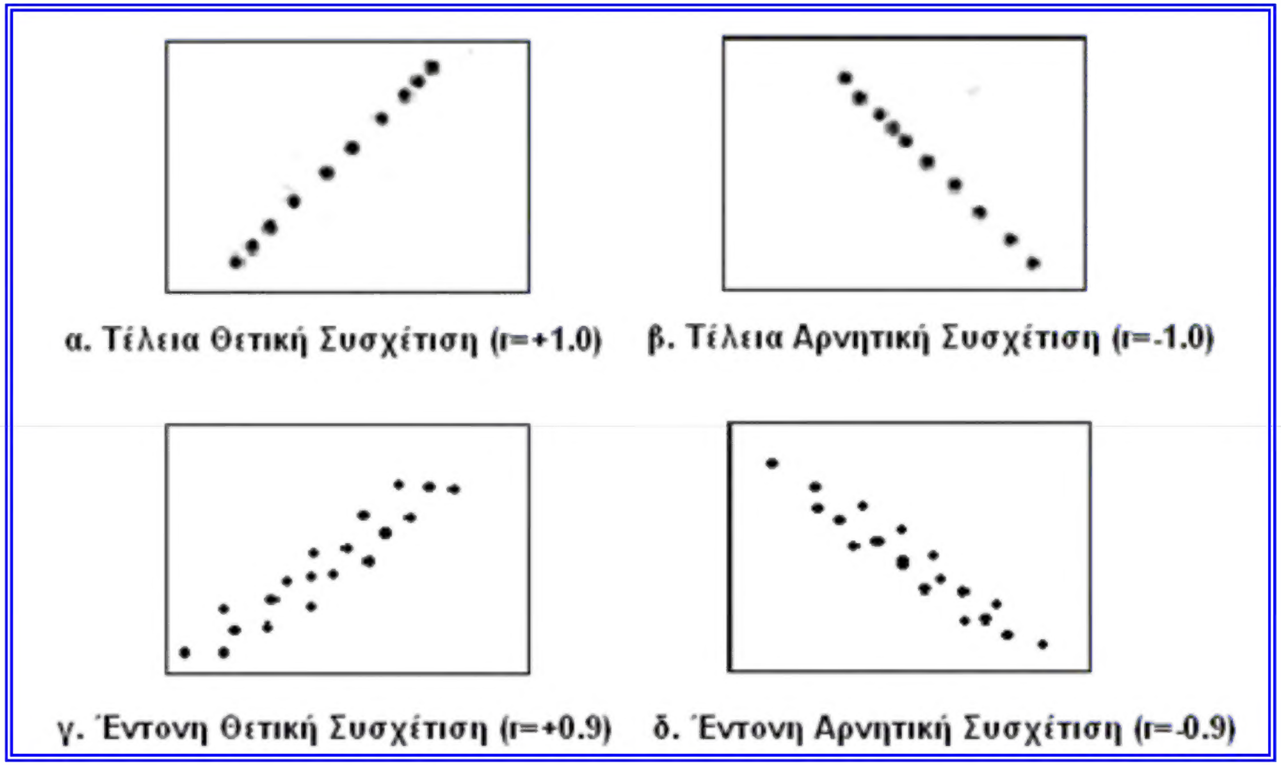
- ✓ **Γραμμική** (linear correlation)
- ✓ **Μη γραμμική** (non linear correlation).

Γραμμική Συσχέτιση (2)

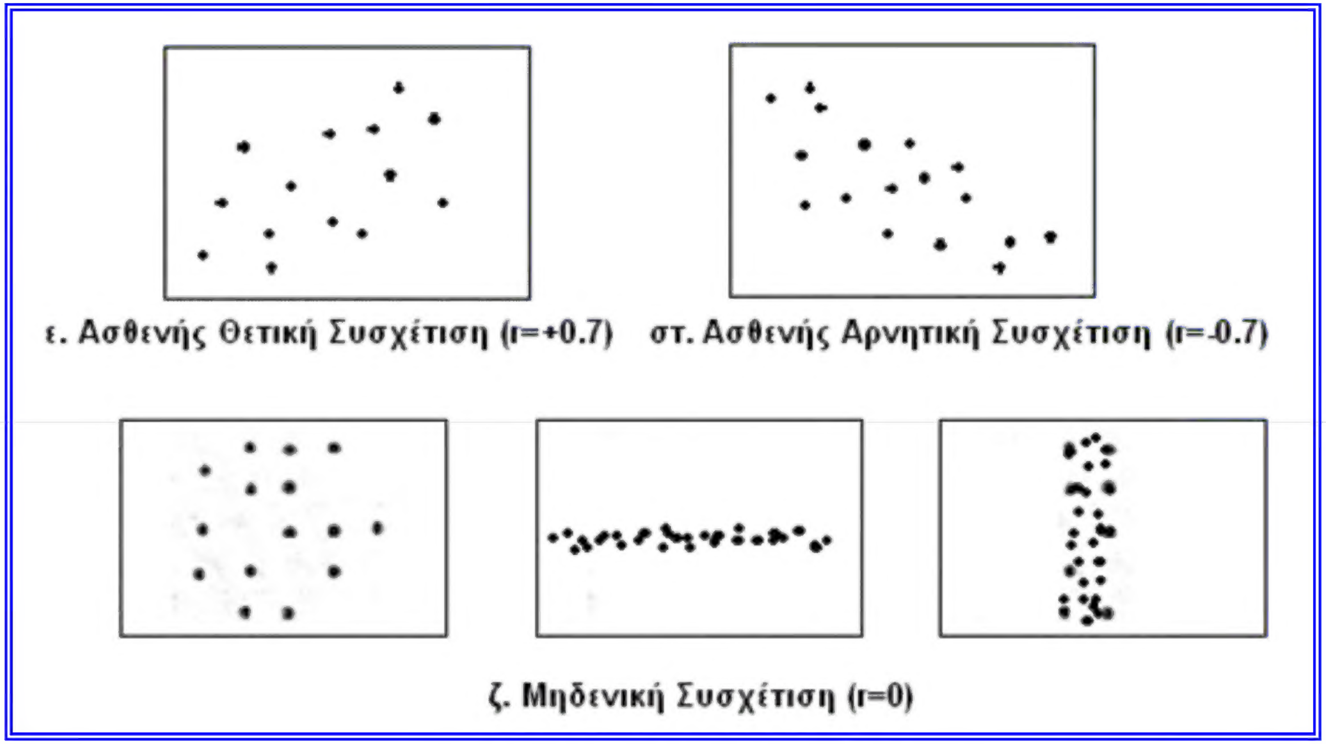
Η συσχέτιση μεταξύ δύο μεταβλητών μπορεί να είναι:

- **Θετική Συσχέτιση** (positive correlation).
- **Αρνητική Συσχέτιση** (negative correlation).
- **Μηδενική Συσχέτιση** (zero correlation).

Γραμμική Συσχέτιση (3)



Γραμμική Συσχέτιση (4)



ε. Ασθενής Θετική Συσχέτιση (r=+0.7) στ. Ασθενής Αρνητική Συσχέτιση (r=-0.7)

ζ. Μηδενική Συσχέτιση (r=0)

Απλή παλινδρόμηση (1)

Η απλούστερη μορφή παλινδρόμησης είναι η απλή γραμμική.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Y_i: η τιμή της εξαρτημένης μεταβλητής

X_i: η τιμή της ανεξάρτητης μεταβλητής

α: το σημείο τομής του άξονα της Υ από την ευθεία παλινδρόμησης

β: η κλίση της ευθείας παλινδρόμησης

Απλή παλινδρόμηση (2)

Κάνουμε λοιπόν τις εξής υποθέσεις:

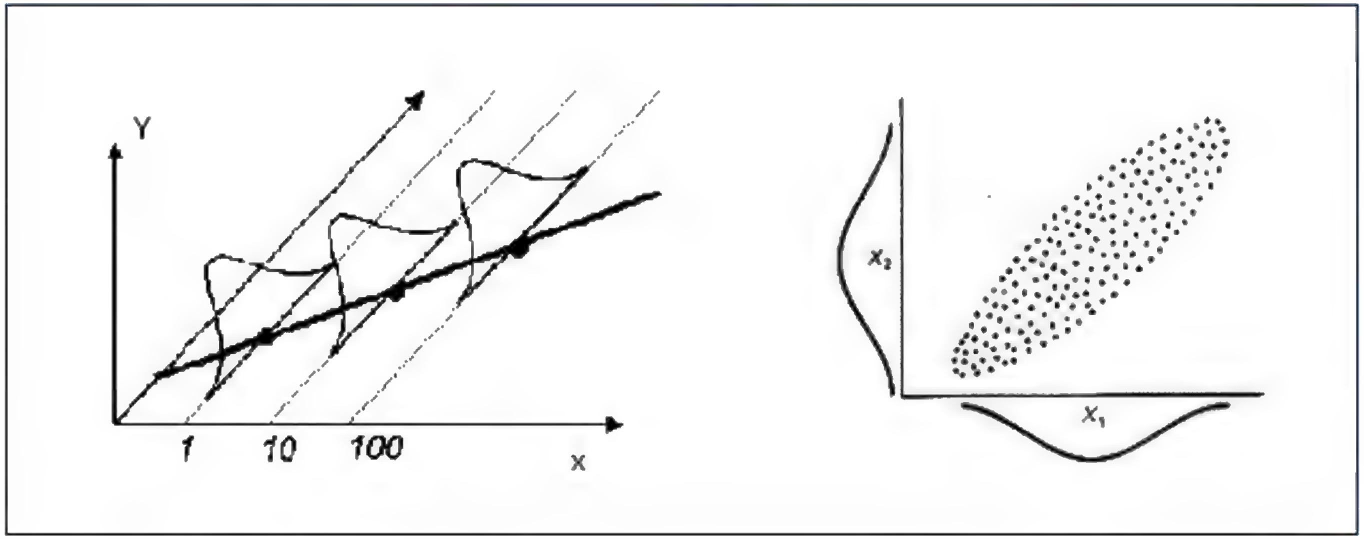
- ✓ Τα $\varepsilon_i \sim N(0, \sigma^2)$
- ✓ Τα ε_i είναι ανεξάρτητα μεταξύ τους

ή ισοδύναμα:

- ✓ Τα $Y_i \sim N(\alpha + \beta X_i, \sigma^2)$, i=1,2…
- ✓ Τα Y_i είναι ανεξάρτητα μεταξύ τους

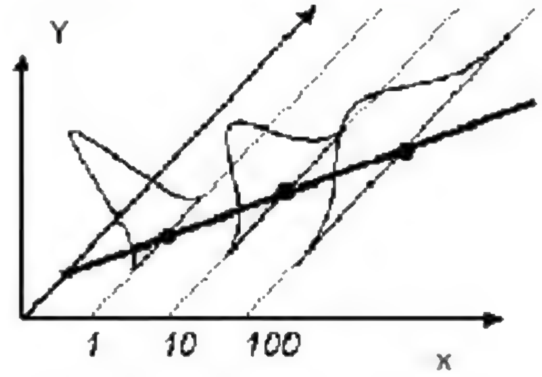
Απλή παλινδρόμηση (3)

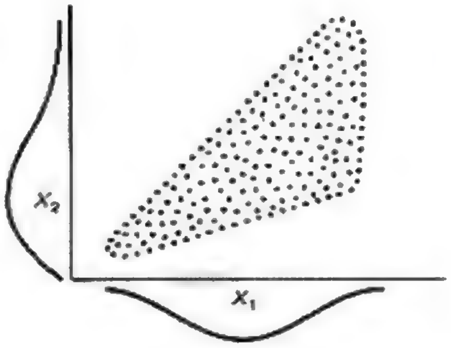
Στα παρακάτω γραφήματα φαίνονται περιπτώσεις ομοσκεδαστικότητας.



Απλή παλινδρόμηση (4)

Στα παρακάτω γραφήματα φαίνονται περιπτώσεις ετεροσκεδαστικότητας.





Σε περίπτωση ετεροσκεδαστικών δεδομένων πρέπει να προχωρήσουμε σε μετασχηματισμό των δεδομένων ώστε να προχωρήσουμε με την εφαρμογή του μοντέλου παλινδρόμησης.

Πολλαπλή Παλινδρόμηση (1)

Το μοντέλο πολλαπλής παλινδρόμησης είναι:

$$Y_i = a + \beta X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i \quad i = 1,2\ldots n$$

Υποθέσεις όσον αφορά τα κατάλοιπα:

- Τα κατάλοιπα ε_i είναι ανεξάρτητα μεταξύ τους και κατανέμονται κανονικά.
- Οι αναμενόμενες μέσες τιμές των σφαλμάτων είναι μηδέν.
- Τα σφάλματα έχουν την ίδια διακύμανση για όλους τους συνδυασμούς των τιμών των ανεξάρτητων μεταβλητών.

Πολλαπλή Παλινδρόμηση (2)

Για να φτάσουμε στο σημείο να επιλέξουμε το κατάλληλο μοντέλο πρέπει να εντοπίσουμε ποιες μεταβλητές πρέπει να μπουν στο μοντέλο και ποιες πρέπει να αγνοηθούν.

Χρησιμοποιούμε τις εξής μεθόδους:

- Backward Elimination
- Forward Procedure
- Stepwise Regression

Πολλαπλή Παλινδρόμηση (3)

Backward Elimination

- ✓ Αρχικά στο μοντέλο εισέρχονται όλες οι μεταβλητές και σε κάθε βήμα αποκλείεται μία μέχρι να απομείνουν εκείνες που συνεισφέρουν στο μοντέλο.
- ✓ Σε κάθε βήμα αποκλείει τη μεταβλητή εκείνη που έχει το υψηλότερο p-value με την προϋπόθεση ότι αυτό είναι μεγαλύτερο από το επίπεδο σημαντικότητας α που έχουμε προκαθορίσει (συνήθως 0.05).

Πολλαπλή Παλινδρόμηση (4)

Forward Procedure

- ✓ Το μοντέλο δεν ξεκινάει με καμία μεταβλητή και στην πορεία προστίθενται εκείνες μόνο που συνεισφέρουν στο μοντέλο.
- ✓ Από το σύνολο των ανεξάρτητων επιλέγεται εκείνη που έχει μεγαλύτερο συντελεστή συσχέτισης με την εξαρτημένη Υ.
- ✓ Στη συνέχεια επιλέγεται εκείνη με τον αμέσως μεγαλύτερο συντελεστή μερικής συσχέτισης με την Υ ενώ παράλληλα διατηρούνται στο μοντέλο και οι προηγούμενες επιλεγμένες μεταβλητές.

Πολλαπλή Παλινδρόμηση (5)

Stepwise Regression

- ✓ *Παρόμοια με την Forward*
- ✓ Σε κάθε βήμα ελέγχεται η υπόθεση ότι η παράμετρος $\beta_j = 0$. Έτσι αν σε κάποιο βήμα προέκυψε κάποια μεταβλητή να γίνει ασήμαντη (δηλαδή να έχει p-value>0.05) τότε αυτή αφαιρείται από το μοντέλο.

Συνδιακύμανση (1)

Έστω ότι έχουμε δύο μεταβλητές που ακολουθούν μία από κοινού κατανομή με μέσες τιμές E (X)=μ_X και E (Y)=μ_Y. Ορίζουμε ως **συνδιακύμανση** (covariance) την εξής ποσότητα:

$$Cov(X,Y) = E(X_i - \mu_X)(Y_i - \mu_Y) \Rightarrow Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Λαμβάνει θετικές και αρνητικές τιμές.

Συνδιακύμανση (2)

- Cov (X, Y) > 0 τότε μιλάμε για θετική σχέση μεταξύ X και Y
- Cov (X, Y) < 0 τότε μιλάμε για αρνητική σχέση μεταξύ X και Y
- Cov (X, Y) = 0 τότε μιλάμε για έλλειψη γραμμικής σχέσης μεταξύ X και Y